Fermi National Accelerator Laboratory

FERMILAB-Conf-98/368

Fermilab Central Mass Storage System as a Test Bed for HPSS

Krzysztof Genser et al.

Fermi National Accelerator Laboratory
P.O. Box 500, Batavia, Illinois 60510

December 1998

Published Proceedings of *International Conference on Computing in High Energy Physics (CHEP '98)*, Chicago, Illinois, August 31 - September 4, 1998

Operated by Universities Research Association Inc. under Contract No. DE-AC02-76CH03000 with the United States Department of Energy

Disclaimer

Distribution

Copyright Notification

Fermilab Central Mass Storage System as a test bed for HPSS

Krzysztof Genser, Alexander Moibenko,
Don Petravick, David Sachs, Joseph Syu

Fermi National Accelerator Laboratory, Batavia, Illinois, USA.

Fermilab Central Mass Storage System (FCMSS) provides mass storage related services to 16 groups of users, mainly Fermilab Fixed Target Experiments. FCMSS is also used as a test bed for High Performance Storage System (HPSS). We report on our more than a year and a half long production experience of using HPSS and show results of various performance tests.

Fermilab Central Mass Storage System (FCMSS) is used by several experimental groups to store and retrieve large amount of physics data. It consists of, IBM 3494 Automatic Tape Library with IBM 3590 tape drives and several IBM RS/6000 machines. Back in 1996 NSL-Unitree was used as FCMSS hierarchical management system (HSM).

Fermilab is participating in HPSS [3] Collaboration as an early deployment site and had built a first HPSS production system in late 1996 using HPSS release 3.1. The system consisted of two IBM RS/6000-580 (580) machines running HPSS Mover Servers and IBM RS/6000-59H (59H) machine running all other HPSS servers. Two major groups were migrated from NSL-Unitree into that system by February 1997 and at that time it contained about 4TB of data in 300000 files [1]. Subsequently, the following upgrades and changes were made: several important AIX and Encina patches were applied, IBM RS/6000-59H machine was upgraded to RS/6000-J50 (J50) machine with 768MB of RAM and two 166MHz processors, HPSS release 3.1 was upgraded to HPSS release 3.2 and a third

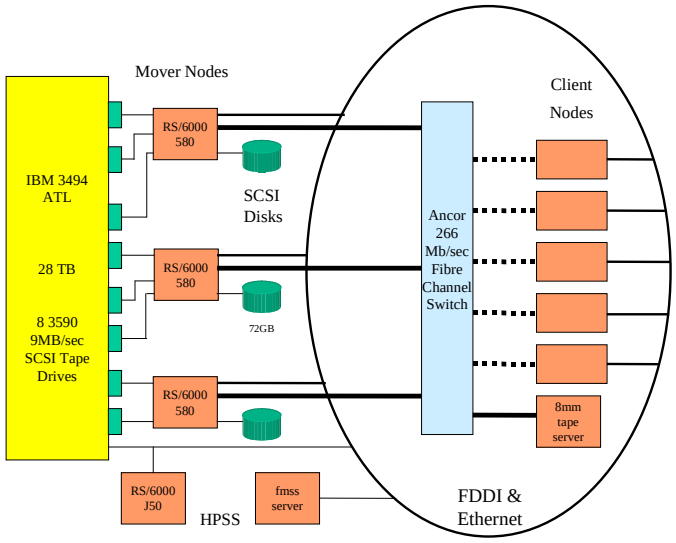

Figure 1: Fermilab Central Mass Storage Configuration as used in the first HPSS I/O Load Test. Current production setup has the 580 replaced with RS/6000-F50.

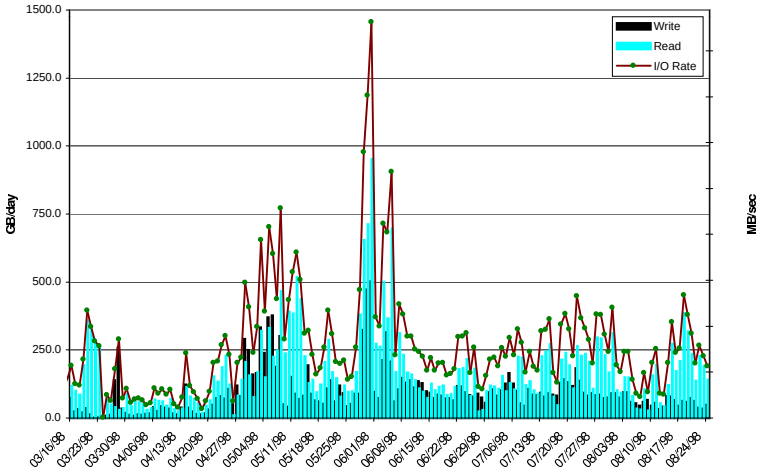

Figure 2: Fermilab Central Mass Storage I/O Usage. Normal user activity is about 300GB/day. The peaks represent various load tests. Maximum sustained load test rate was 23.4MB/sec.

580 machine was added as a HPSS Mover Server. All 580 machines were equipped with 512MB of RAM and about 70 GB of disk space each by early 1998. Final NSL-Unitree to HPSS migration occurred in March 1998. At that time the system contained 18.3TB of data in 355000 files.

Figure 1 shows CFMSS configuration after decommissioning NSL-Unitree. Each of the 580 machines had a separate FDDI ring connection to a DEC FDDI GIGAswitch and a Quarter Speed FibreChannel (FCS) connection to an Ancor FibreChannel switch. Current (August 1998) CFMSS configuration is similar to the one shown, with 580 machines replaced by RS/6000-F50 (F50) machines with 512MB of RAM and two 166MHz processors.

FCMSS users access their data via a UNIX shell like interface called Fermilab Mass Storage System (fmss) [1]. The interface was created to hide the underlying HSM, provide HPSS access from non-DCE nodes, provide retry and recovery features, implement quota and data grouping, make usage of the HPSS pftp interface more transparent and to be able to limit the number of concurrent HSM accesses. Two other HSM user interfaces are provided to export HSM data to 8mm tapes and to cache 8mm tape data into the HSM.

fmss is implemented as a client server system. The fmss client provides the following subcommands: *cp*, *rm*, *mkdir*, *mv*, *ls*, *chmod* which have a very similar functionality as their UNIX counterparts. Other subcommands are fmss specific, e.g. *stage* moves files from HSM tape to HSM disk, *query* informs whether the object is a file and if it resides on HSM tape or disk, *wait* in combination with *cp* allows for waiting for a file transfer completion when retrieving files from HSM. The remaining subcommands *version*, *status* and *activity* display fmss version, system availability and active fmss transactions. The following is an example of copying of a local disk file into HSM using fmss: *fmss cp disk:./localfile mss:experiment/directory/fileinhsm*

In August 1998, fmss was used by 16 groups of users who stored about 23TB of data in approximately 440000 files. File size ranged from a few kilobytes to above two Gigabytes. The above data was stored on 2630 tapes.

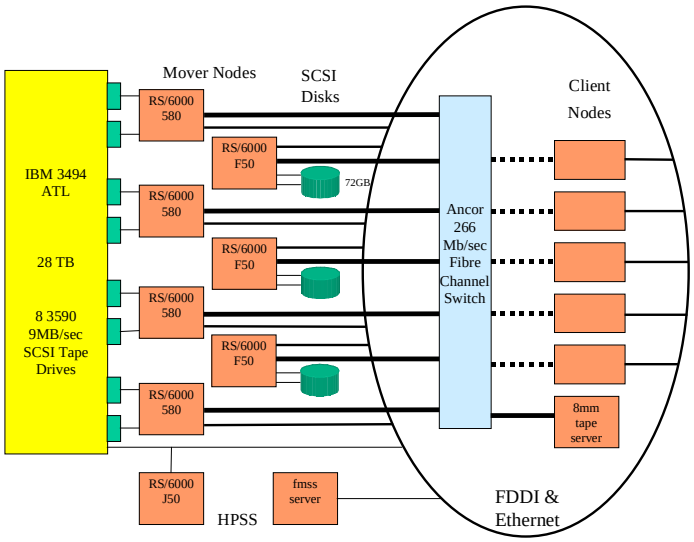

Figure 3: Fermilab Central Mass Storage Configuration in the Second HPSS I/O Load Test

Currently FCMSS adopts cache like space policy. As users remove their files from the system, the tapes are repacked and reclaimed each night to provide space for new files. Figure 2 shows fmss usage in GB transferred per day. Normal user activity is about 3MB/sec daily average. The peaks on the plot represent days when load tests were performed.

FCMSS HPSS/fmss is configured in the following way: Each fmss group has its own HPSS Class of Service, all fmss Classes of Service map to Disk Over Tape HPSS Storage Hierarchies. One HPSS Disk Storage Class (of about 130GB now) is shared among all fmss user groups and is served by two F50 machines. This allows efficient use of the disk space as the activity of different groups rarely peak at the same time. Each fmss user group has its own Tape Storage Class to make sure that data belonging to one experiment is not mixed with data from another experiment. Due to different access patterns and shorter caching time, of the order of ten days, another Disk Storage Class (of about 70GB) together with another Tape Storage Class is used for 8mm tape caching and is served by another F50 machine. To match the speed of the tapes with the disk speed, disk Storage Classes are configured to be 4way HPSS stripes; no HPSS tape striping is used. To make sure that migration to tapes can keep up with the speed of the network connections, disk to tape migration processes can allocate up to three tape drives at a time. To prevent scattering of data over many tapes a maximum of 4 tapes can be open for writing for any Tape Storage Class at any given time.

To make sure that there are enough resources available to satisfy active requests fmss is configured to allow for 40 concurrent transactions plus 10 8mm tape caching jobs. HPSS NFS access is almost disabled to be able to limit the total number of transfers. To optimize read access, lists of files on HPSS tapes are produced weekly. This information is used to group fmss file stage requests to minimize the number of tape mounts.

In order to be able to estimate the hardware resources needed to provide certain I/O capability several tests were performed.

First, component I/O tests were done and the following results were obtained: Quarter Speed FibreChannel (FCS) interface bandwidth was found to be about 22MB/sec, memory bus band-

Table 1: Results of the first and second HPSS I/O Load Test

	First Test with 3×580			Second Test with $3 \times$F50+4×580		
	I/O Volume in GB	Number of files	I/O Rate in MB/sec	I/O Volume in GB	Number of files	I/O Rate in MB/sec
Read by load test	122	122	5.8	219	219	15.6
Write by load test	90	90	4.3	96	96	6.8
Read by users	47	190	2.2	3	30	0.2
Write by users	21	134	1.0	10	119	0.7
Total	280	536	13.3	329	464	23.4

width [4] was measured to be 280MB/sec for RS/6000-580 machines and 160MB/sec/processor for RS/6000-F50 machines (with 166MHz CPU), cache disk bandwidth was 5MB/sec (2/3rd of disks) and 10MB/sec (1/3rd of disks), SCSI adapter bandwidth was 17.7MB/sec for 580s and 20.0MB/sec for F50s.

Second, two I/O load tests were performed. Figure 1 shows CFMSS configuration in the first I/O test, where three RS/6000-580 machines were used as HPSS Mover Servers. During the test 1GB uncompressible files were transferred. The data was transferred to and from the external disks via HPSS disk cache from and to HPSS tapes. There were two writing and four sequential reading processes, all using FibreChannel network. The reading processes used the tape list information to group read requests by tapes to minimize the number of tape mounts. The test lasted for six hours and the average sustained transfer rate was 13.3MB/sec. In this test CPU power of the 580s was the main limiting factor. Table 1 shows the details on the amount of data transferred during the first I/O test. Figure 3 shows CFMSS configuration in the second I/O test. Here four RS/6000-580 machines were used as HPSS Tape Mover Servers and three dual processors RS/6000-F50 machines were used as HPSS Disk Mover Servers. This time there were two writing and eight sequential reading processes. One writing and three reading processes used FDDI network with a combined bandwidth of 12MB/sec. The remaining I/O processes used FCS network. This test lasted for four hours and the average sustained transfer rate was 23.4MB/sec. In this test the limiting factors were: bandwidth of FibreChannel interfaces on F50s (22MB/sec), disk I/O bandwidth on two of the F50s (5MB/sec) and migration from HPSS disk to tape (5MB/sec, this rate was then made to be 8MB/sec after the test). Table 1 shows the details on the amount of data transferred during this test.

Third, CPU/Byte of the moved data was measured. Table 2 shows the numbers. The table indicates that local tape to/from disk data transfers are about 9 times more CPU efficient as compared to the network transfers (e.g. item 1 vs. 3+6).

The above tests enable one design a configuration of an HPSS Mover machine (Scalable HPSS Mover Unit) which can be then multiplied to achieve a desired I/O capability. The following Unit should be able to sustain an average transfer rate of about 15MB/sec (for 1GB files), be able to run its tape drives at their nominal speed and provide a disk buffering equivalent to 3 hours of its tape drive transfers. The Unit should consist of 1 RS/6000-F50 machine with 4 166MHz processors and 512MB of RAM with 7 SCSI adapters (7*20MB/sec), 3 IBM3590 tape drives, each on a separate SCSI adapter (3*9MB/sec=27MB/sec), 32 9GB 10MB/sec SCSI disk drives which should be 4-way striped across 4 SCSI adapters (4*10MB/sec=40MB/sec) and a 125MB/sec capable network adapter.

Based on Fermilab Tevatron Run II Committee decision HPSS is a backup HSM solution for Tevatron Run II. Enstore[2] is under development as the main HSM. HPSS will continue to be used at Fermilab as HSM product for next Fixed Target run in 1999 and when required to collaborate

Table 2: RS/6000 F50 2 × 166MHz Processor CPU required to move 1GB of data.

	RS/6000 F50 except as noted 580 Average CPU for two F50 processors	CPUsec for 1GB	MB/CPUsec	Approx Single Transfer Rate MB/sec
1	HPSS Disk to HPSS Local Tape	4.9 ± 0.4	209.5 ± 18.5	8
2	HPSS Tape to HPSS Local Disk	4.9 ± 0.8	210.8 ± 35.6	8
3	HPSS Disk to FCS	20.6 ± 1.1	49.6 ± 2.6	8
4	FCS to HPSS Disk	21.3 ± 0.6	48.0 ± 1.3	8
5	HPSS Tape to FCS	21.2 ± 0.1	48.3 ± 0.3	8
6	FCS to HPSS Tape	21.7 ± 0.3	47.2 ± 0.6	8
7	pftp FCS to HPSS Disk	21.6 ± 0.3	47.3 ± 0.7	18 Max
8	pftp HPSS Disk to FCS	24.8 ± 0.6	41.3 ± 1.0	22 Max
9	FCS to HPSS Tape (580)	75.0 ± 1.6	13.7 ± 0.3	8
10	HPSS Tape to FCS (580)	65.0 ± 2.3	15.8 ± 0.6	8

with other institutions.

In order to meet the storage demands of Fixed Target 1999 run FCMSS will undergo several changes. HPSS disk cache will be upgraded to a total of 96 9GB 10MB/sec disks. Internal HPSS network will be upgraded to a Gigabit/sec level. Hardware RAID will be used to replace current mirrored SCSI disks to store HPSS metadata. Operations will be automated as much as possible using non-GUI procedures and scripts to start, monitor, restart and shutdown HPSS, which should allow to minimize operational effort and allow for non-expert off-hours operation.

In summary: Fermilab Central Mass Storage System is fully HPSS based since NSL-Unitree to HPSS migration which was done in March 1998. Several HPSS I/O tests were performed, a sustained aggregate rate of 23.3MB/sec was achieved. CPU per Byte moved within and to and from HPSS was measured. HPSS was declared to be a secondary MSS solution for Tevatron Run II. Enstore product is under development.

References

[1] K. Fidler, K. Genser, S. Kalisz, J. Mack, A. Moibenko, D. Sachs *Mass Storage Systems at Fermilab: An early experience with the High Performance Storage System, International Conference on Computing in High Energy Physics*, Berlin, April 1997.

[2] D. Petravick at al. *ENSTORE - an Alternate Data Storage System*, these proceedings.

[3] D. Teaff, D. Watson and R.A. Coyne, *The Architecture of the High Performance Storage System, Fourth NASA Goddard Conference on Mass Storage Systems Technologies*, March 1995.

[4] using stream copy as defined in http://www.cs.virginia.edu/stream.

[5] IEEE Storage system Standards Working Group (SSSWG) (Project 1244), *Reference Model for Open Storage Systems Interconnection, Mass Storage Reference Model Version 5*, Sept.1994.

[6] Open Software Foundation, *Distributed Computing Environment Version 1.0 Documentation Set*. Open Software Foundation, Cambridge, Mass. 1992

[7] Dietzen, Scott, Transarc Corporation, *Distributed Transaction Processing with Encina and the OSF/DCE*, Sept. 1992.